NPCMJ への PropBank 形式の意味役割と概念フレームの付与の進捗報告

竹内 孔一
岡山大学大学院
takeuc-k@okayama-u.ac.jp

バトラー アラステア
弘前大学

長崎 郁
名古屋大学

パルデシ プラシャント
国立国語研究所
prashant@ninjal.ac.jp

1 はじめに

Web 上で閲覧可能な日本語統語解析情報付きコーパス (ツリーバンク)NPMCJ(NINJAL Parsed Corpus of Modern Japanese) [1] が国立国語研究所で開発が進んでおり既に構文木が公開されて言語学者や言語学習者に利用されている[1)]．現在 14 ジャンルの出典で約 4 万文の構文木が構築されておりユーザはツールを通して事例を確認できる．本研究では NPCMJ に対して述語項構造シソーラス[2)] の概念フレームをベースとして名前による意味役割 (「動作主」,「対象」など) と PropBank 形式の意味役割 (Arg0, ArgM-ADV など) を付与している [2]．例えば「両親は，長女の愛は厳しく育てた」の場合，概念フレームと意味役割は下記のように付与する．付与過程で英語の構文には出てこない自動

図 1 「両親は，長女の愛は厳しく育てた」の概念フレームと意味役割付与例

詞の受動態について PropBank では議論されてこなかった必須項が必要であることを先行研究で発表した [3]．NPCMJ は現在も付与が続いているコーパスで 6 万文まで拡張されることが計画されており，本研究も継続して意味役割と概念フレームを付与する予定である．本論文では付与作業と現段階での付与された意味タグの内容について記述する．

1） http://npcmj.ninjal.ac.jp/
2） http://pth.cl.cs.okayama-u.ac.jp/

2 NPCMJ 内の付与対象の設定

NPCMJ は NP や PP といった非終端記号，SBJ や OBJ などの区別，トレース，参照情報が付与されている．一方で，意味役割を付与する述語と項のみをタグ付けした情報は存在しないが，付与されている詳細な構文情報から論理形式の意味構造に変換する手法が提案されており [4]，これを基に付与すべき述語と項が設定される．

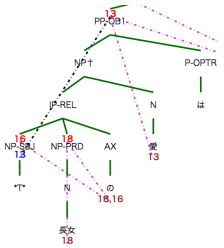
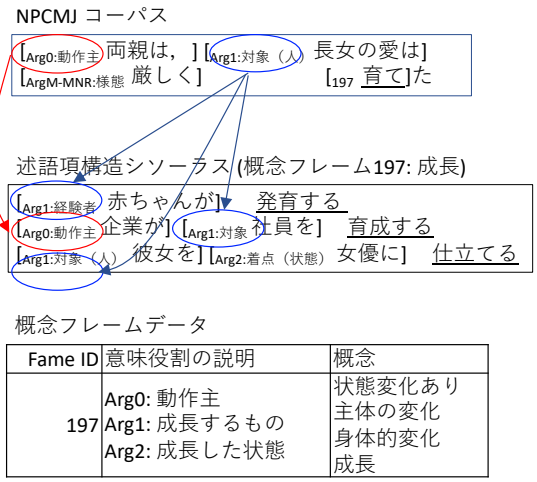


図 2 「長女の愛は」に関する NPCMJ の構文木の例

3 概念フレームデータに基づく PropBank 形式の意味役割

意味役割とは文における述語の係り関係を想定した場合，述語と項との意味的な関係を記述するタグである．例えば図 1 の場合，述語「育てる」に対して，係り元は「両親は,」「長女の愛は」「厳しく」の 3 つである．このうち，「厳しく」には述語に対する修飾要素として ArgM-MNR を付与し，残り二つは項として Arg0 と Arg1 を付与する．

PropBank では意味役割は述語の各語義ごと (つまり述語に対してより細分化された分類) に付与されているが，本研究では述語項構造シソーラスの概念フレームを利用して概念フレームごと (複数の述語で共通する概念) に対して



Fame ID	意味役割の説明	概念
197	Arg0: 動作主 Arg1: 成長するもの Arg2: 成長した状態	状態変化あり 主体の変化 身体的変化 成長

図 3 NPCMJ の付与コーパスと述語項構造シソーラスおよび概念フレームデータの関係

一貫した必須項の意味役割を付与する．つまり概念フレームが同一の述語間では Arg0，Arg1 などの意味的な関係が一貫した意味役割で付与される．これにより，同じフレーム内での言い換え関係などが同じタグで取り出せること，さらに，付与の際に，事例が少なくても概念フレーム内で述語があれば意味役割の設定が可能なためより安定して意味役割を設定できることが期待できる．

具体的な例をあげて説明する．図 3 では「育てる」は述語項構造シソーラスの概念フレーム番号 197 (状態変化あり/主体の変化/身体的変化/成長) に相当する語義であることを指摘しており，この概念フレームには「熟す」「発育する」「飼育する」「仕立てる」など自動詞，他動詞が含まれている．これらの述語で必須項を表す Arg0，Arg1 は述語項構造シソーラスに事例が付与されているのと同時に図 3 に示すように概念フレームデータを作成して一貫性を保っている．

4 付与作業

付与作業者に対して 2 節で生成された述語と係り元が提示される．統語情報から述語と判定されたものが自動で選択されており，1 文で複数の述語と項の関係が取り出されている (図 4 参照)．

作業者は対象事例を基に述語項構造シソーラスの事例を調べて，合致する概念フレームを探す．合致する概念フレームがあれば，必須項を

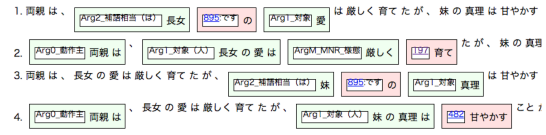


図 4 NPCMJ の付与コーパスと述語項構造シソーラスおよび概念フレームデータの関係

事例から選び，付加詞の場合はあらかじめ決められているリスト [2] から意味役割を選択する．シソーラスに存在しない場合は，該当する概念フレームがある場合は既にある述語を利用して記述し (5 図参照)，必要があれば辞書を更新する．同様に意味役割が不足している場合は辞書管理者が判断して概念フレームの必須項を追加する．

5 FrameNet に準拠したデータ形式

意味役割と概念フレームを付与した NPCMJ からデータを取り出す形式として現在 2 種類が可能である．1 つは図 5 に示すように FrameNet に準拠した XML 形式であり，文と述語および意味役割が付与されている．FrameNet の XML では述語と項や係り元の構造が理解しやすいく簡素であるため 6 節の統計量も XML 形式のデータから取り出している．

しかしながら XML 形式には NPCMJ の統語情報が付与されていない．そこで本論文ではまだ利用していないが構文木の構造による出力形式も構築している．統語情報や参照情報，形態素情報など多量な情報が付与されており付与された全ての情報を取り出すことが可能である．

```
<sentence corpID="26936" docID="4656" sentNo="1" paragNo="1" aPos="0"
   ID="1_misc_1709kytext1">
  <text>愛 と 真理 は 2 歳 違い の 姉妹 * 。</text>
  <annotationSet cDate="01/05/21 16:48:02 JST Tue" status="UNANN"
   ID="13879">
～略～
<annotationSet cDate="01/05/21 16:48:02 JST Tue" luID="0" luName="の.v"
 frameID="895:です" frameName="895:です" status="MANUAL" ID="13880">
 <layer rank="1" name="Target">
  <label end="16" start="16" name="Target"/>
 </layer>
 <layer rank="1" name="FE">
  <label feID="2" bgColor="FFFF00" fgColor="000000" end="19" start="18"
   name="Arg1_対象"/>
  <label feID="1" bgColor="FFFF00" fgColor="000000" end="14" start="9"
   name="Arg2_補語相当（は）"/>
 </layer>
```

図 5 FrameNet に準拠した NPCMJ 意味役割付与データ形式

図 5 ではコーパス misc_1709kytext1 にテキストがあったことがわかる．述語と項の情報は annotationSet のタグ内に記述されてお

り，複数の述語と項の組がある場合は複数の annotationSet が記述される．ここでは「の」が述語で項は「2 歳違い」と「姉妹」である．連体修飾の形になっている．

XML では述語「の」の概念フレームは 895(コピュラ) に登録されている述語「です」に相当することを示している．「2 歳違い」と「姉妹」の意味役割はそれぞれ Arg1:対象 と Arg2:補語相当（は）である．つまり平叙文だと

- [$_{Arg1:対象}$ 姉妹は] [$_{Arg2:補語相当（は）}$ 2 歳違い] [$_{895}$ です]

として付与している．

6 現在の付与内容

現段階で付与している述語の付与数は 22,637 箇所でそのうち項の付与数は 44,404 箇所である．各要素について以下に説明する．

述語について

基本的に NPCMJ では文内の表層形に述語の概念を付与している．よって述語の種類数を求める場合，単純に文字列で集約すると付与述語の語彙数がわからない．そこで形態素解析 MeCab を通して述語の種類を推定して求めた．その結果，付与している述語の種類数は 4048 種類である．形態素解析を利用してどのような種類の述語があったかを図 6 に示す．

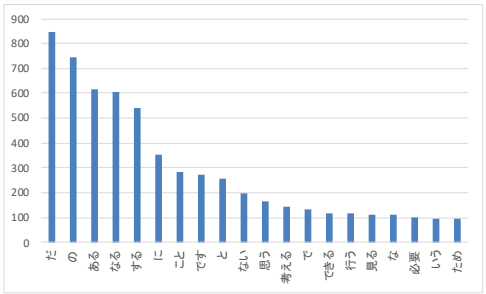


図 6 付与した述語 (上位 20 件)

図 6 は縦軸が形態素の頻度を表している．「だ」がもっとも多く続いて「の」(同格)，「ある」「なる」などの和語動詞が多い．

次に述語の品詞について頻度の多い述語から順に図 7 に上位 20 件まで示す．品詞は MeCab (IPAdic) の品詞細分類までとする．述語の頻度で 3 番目に「名詞 一般」が多いのは「こと」を述語として付与しているためである．「こと」は非飽和名詞 [5, 6] であり，内容を名詞化する

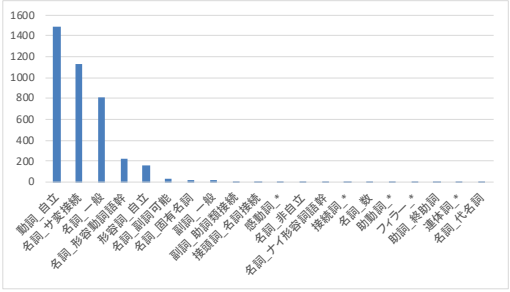


図 7 述語の品詞 (上位 20 件)(MeCab による推定)

際に使われる．述語項構造シソーラスでは現段階では非飽和名詞に関する項目が無いが可能な限り付与している．現段階では一部概念フレームとして WordNet の概念フレームを付与している (例えば「こと」に対して「wn:00034479-n」)．また「話」「仕事」などの非飽和名詞の他に「試み」や「喜び」など述語的な要素を持つものなどがある．これらは述語項構造シソーラスの概念フレームを付与している．よって先行研究 [7] で付与されている事態性名詞についても今回の NPCMJ コーパスに付与されている．

次に概念フレームについて図 8 に上位 10 件の頻度分布を示す．最も多いのが「コピュラ」

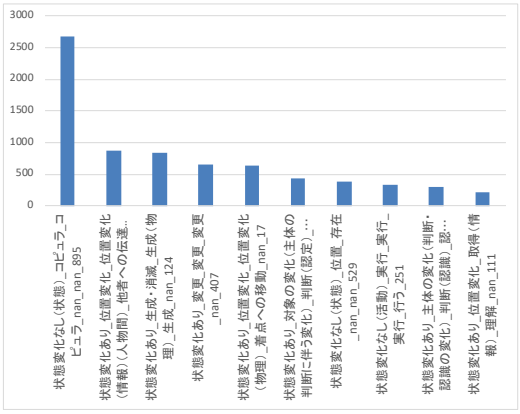


図 8 述語の概念フレーム (上位 10 件)

(ある) である．次に「伝達」(話す)，「生成」(作る)，「変更」(変える)，「移動」(行く) が多いことがわかる．

項について

項に付与した意味役割の統計量について示す．PropBank 形式の意味役割上位 20 件を図 9 に示す．

図 9 では最も多い意味役割は Arg1 で Arg0 の 1.6 倍程度出現している．必須項は Arg2 までが数千の単位で出現していて，割合が多いことが

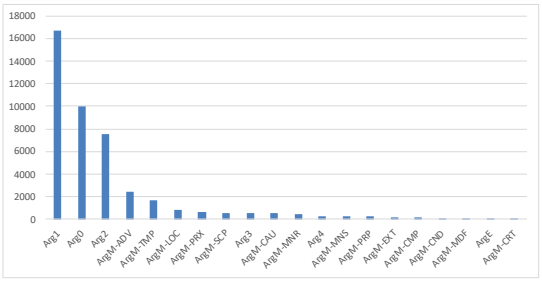


図 9　PropBank 形式の意味役割 (上位 20 件)

分かる．一方，Arg3 や Arg4 はこれらに比べると相対的に少ないことが分かる．

付加詞に対する意味役割では副詞を表す ArgM-ADV (副詞)，ArgM-TMP (時間)，ArgM-LOC (場所) がそれぞれ Arg2 に続いて出現している．ArgM-PRX は連語である．このタグは主動詞の意味がほとんどなく項の部分に中心的な意味がある場合に付与する．「〜をする」という軽動詞構文はこのタイプの意味役割を付与している．

- [$_{ArgM-PRX}$:連語 感動を] [$_{187}$:感動する する] (book_except-15)

また，慣用句などで項の部分と動詞の部分が一つになってある決まった意味を表す場合の項の部分に付与する．

- [$_{ArgM-PRX}$:連語 手の] [$_{547}$:手間取る かかる] (aozora_Tsuboi-1968)

よって，ArgM-PRX を付与している項は述語の一部と見なしている．

次に名前の意味役割上位 20 件の頻度分布を図 10 に示す．「対象」が最も多く，次いで「動作主」「経験者」が続く．意味役割の頻度分布が PropBank 形式の意味役割と異なるのは Arg1 は「対象」や「経験者」など複数の名前の意味役割に付与されることが原因である．これは PropBank 形式の意味役割が構文を重視した分類であるのに対して，名前意味役割は意味的な関係を重視している点に起因する．

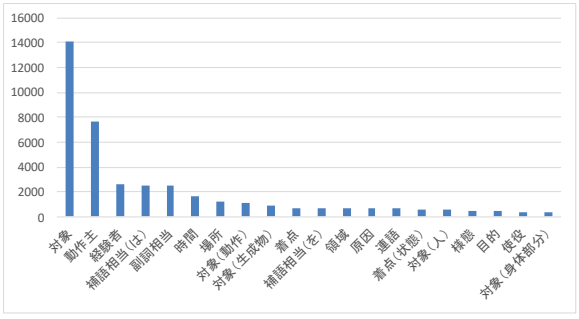


図 10　名前の意味役割 (上位 20 件)

7　最後に

本論文では現在進めている NPCMJ に対する概念フレームと意味役割付与データ構築プロジェクトでどのようなタグをどの程度付与しているかについて記述した．PropBank 形式の意味役割を採用し，概念フレームごとに意味役割の定義集合を一貫して構築している．NPCMJ における統語情報付与作業はこれからも続くため，今後も意味役割と概念フレームの付与を続けると同時に付与データを公開していく予定である．

8　謝辞

本研究の遂行にあたって国立国語研究所機関拠点型基幹研究プロジェクト「統語・意味解析コーパスの開発と言語研究」および科研費（課題番号 19K00552) の助成を受けた．

参考文献

[1] 吉本啓, 周振, 小菅智也, 大友瑠璃子, Alastair Butler. 日本語ツリーバンクのアノテーション方針. 言語処理学会第 19 回年次大会, 2013.

[2] 竹内孔一, バトラーアラステア, 長崎郁, ホーン スティーブンライト. PropBank スタイルの意味役割タグを導入した述語項構造シソーラスと NPCMJ への付与計画. 言語処理学会第 25 回年次大会, pp. 136–138, 2019.

[3] Koichi Takeuchi, Alastair Butler, Iku Nagasaki, Takuya Okamura, and Prashant Pardeshi. Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC2020)*, 2020.

[4] Alastair Butler. From discourse to logic with stanford corenlp and treebank semantics. In *New Frontiers in Artificial Intelligence*. JSAI-isAI 2019, vol. 12331 of Lecture Notes in Computer Science, 2020.

[5] 西山佑司. 日本語名詞句の意味論と語用論. ひつじ書房, 2003.

[6] 影山太郎. 日英対照 名詞の意味と構文. 大修館書店, 2011.

[7] 小町守, 飯田龍, 乾健太郎, 松本裕治. 名詞句の語彙統語パターンを用いた事態性名詞の項構造解析. 自然言語処理, Vol. 17, No. 1, pp. 141–159, 2010.